OPAL
OPEN DATA PORTAL

Deliverable D7.1
Suchkomponente

Autoren: Adrian Wilke
Reviewer: Caglar Demir

Veröffentlichung	Öffentlich
Fälligkeitsdatum	30.06.2019
Fertigstellung	02.07.2019 finale Prüfung 28.06.2019 v1.0
Arbeitspaket	AP7
Typ	Softwaredokumentation
Status	Final
Version	1.0

Kurzfassung:
Dieses Deliverables beschreibt die Verbesserung der Auffindbarkeit von Datensätzen durch die Erweiterung der Suchfunktion um Synonymlisten. Dazu werden linguistische Wissengraphen betrachtet und die Extraktion von Synonymlisten beschrieben. Zur Indexierung von Texten und Einbindung der Synonyme findet der Suchserver Elasticsearch Verwendung.

Der für dieses Deliverable verwendete Code ist unter [OPAL-LA] verfügbar.

Schlagworte:
Suche, Synonyme, Elasticsearch, DBnary, lemon

mFUND
Das Startkapital für die Mobilität 4.0

Bundesministerium
für Verkehr und
digitale Infrastruktur

Inhalt

1 Einführung Suchkomponente

Über die OPAL **Benutzerschnittstelle** haben Anwender die Möglichkeit, nach beliebigen Begriffen in Metadaten zu **suchen**. Eine Anforderung an das System ist die schnelle Generierung einer passenden und gewichteten Ergebnisliste. Zur Generierung werden sowohl die Datenstrukturen der Metadaten als auch die entsprechenden Texte der derzeit in OPAL vorhandenen rund **800.000 Datensätze** einbezogen.

Im OPAL Arbeitspaket zu Indexstrukturen ist die Einbindung des Suchservers **Elasticsearch** [Elasticsearch] beschrieben (siehe Deliverable D4.3). Durch die damit vorgelagerte **Indexierung** wird eine performante Suche umgesetzt, die in der hier beschriebenen Suchkomponente verwendet wird.

Auch mit der Indexierung bleibt das **Problem** der Verwendung von Suchbegriffen bestehen, die nicht in den ursprünglichen Metadaten enthalten sind. Obwohl ein Anwender semantisch korrekte Begriffe verwendet, werden potentiell verfügbaren Datensätze nicht gefunden. Beispielsweise könnte ein Volltext das Wort “Stadtbahn” enthalten. Die Verwendung von Suchbegriffen wie “S-Bahn” oder “Straßenbahn” würde den entsprechenden Datensatz ggf. nicht finden. Um die **Auffindbarkeit** zu verbessern können **Synonyme** (unterschiedliche Begriffe ähnlicher Bedeutung) aufgelöst werden. Hierdurch werden zusätzliche semantisch verwandte Begriffe in die Suche einbezogen.

2 Synonyme in Wissensgraphen

Wörterbücher und **Lexika** stellen vermehrt Forschungsgegenstände in den Domänen Semantic Web und Linked Data dar. Dabei werden bereits verfügbare offene Ressourcen häufig verwendet, um die entsprechenden linguistischen Daten zu klassifizieren, als semantische Daten auszuzeichnen und miteinander zu verlinken. Die folgende Abbildung zeigt einen Überblick über häufig verwendete linguistische Wissensgraphen [LingHub]:

Abbildung: Linguistic Linked Open Data Cloud [LingHub]
Bildquelle: LIDER http://linghub.org/llod-cloud (Daten unter CC BY-NC-SA 4.0)

Die einzelnen linguistischen Wissensgraphen werden typischerweise initial mit einem Schwerpunkt erstellt (z.B. ein Fokus auf eine Sprache oder eine bereits verfügbare Datenquelle) und beinhalten daher jeweils nur einen Teil der potenziell verfügbaren Daten. Zur Anwendung in OPAL werden **Synonyme deutschsprachiger Nomen** benötigt. Die folgende Tabelle zeigt eine verfeinerte Auswahl von Wissensgraphen, die Lexika oder Wörterbücher enthalten:

Name	Referenz	Notizen
BabelNet	[BabelNet]	Limitierung der Anzahl von API Zugriffen.
DBnary	[DBnary]	Mehrsprachig. Creative Commons Attribution-ShareAlike 3.0.
RDF/OWL Representation of WordNet	[WordNet-RDF]	Datensatz umfasst ausschließlich Amerikanisches Englisch.

Tabelle:: Lexika und Wörterbücher als Wissensgraphen (Auswahl)

Das bekannteste Datenbank für semantische und lexikalische Beziehungen ist **WordNet**. Die dort enthaltenen Begriffe sind als RDF-Repräsentation/Wissensgraph verfügbar. Da die Datenbank ausschließlich Begriffe in amerikanischen Englisch umfasst, genügt sie den OPAL Anforderungen nicht. Eine weiterer verbreiteter Wissensgraph ist **BabelNet**. Dieser eignet sich generell zur Einbindung in OPAL. Lediglich die Limitierung der Anzahl der API-Zugriffe spricht der Einbindung in OPAL entgegen. (Für Forschungszwecke kann jedoch eine Erhöhung des Limits beantragt werden.) Die finale Wahl des einzubindenden Wissensgraphen fiel auf das frei verfügbare **DBnary**. Dieser Graph wurde aus Wiktionary Daten erstellt [DBnary15, Wiktionary] und enthält Informationen über Synonyme mehrerer Sprachen.

2.1 Extraktion von Synonymen

In DBnary finden aufeinander aufbauende **Ontologien** Verwendung. Dies sind *Lexicon model for ontologies (lemon)* [lemon, lemonw3], *LexInfo* [LexInfo] und eine Erweiterung für *DBnary* [DBnary] selbst. Die folgende Abbildung zeigt die zur Extraktion relevanten Ressourcen und Beziehungen:

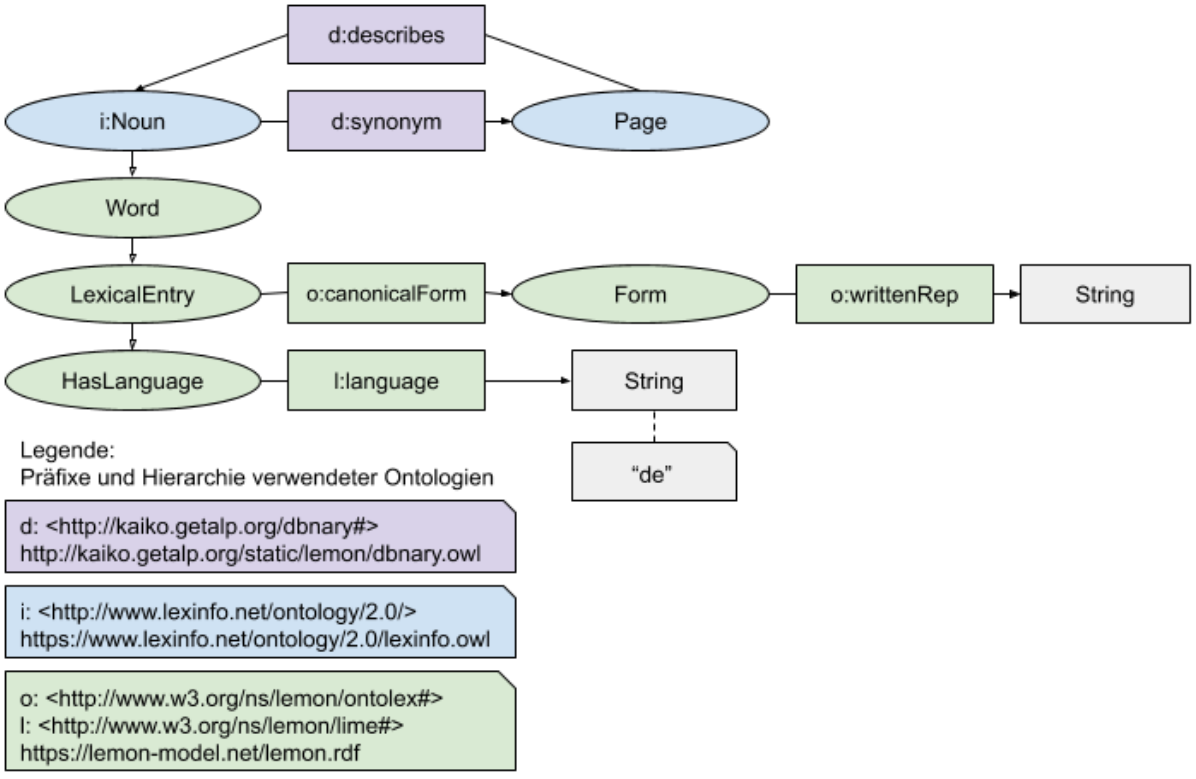

Abbildung: Ontologie zur Extraktion linguistischer Ressourcen

Die Extraktion geschieht über eine SPARQL Anfrage und folgender **Verfeinerung**:

1. Filterung aller Nomen in deutscher Sprache,
2. Einschränkung auf die deutschsprachigen Nomen, für die Synonyme spezifiziert sind,
3. Extraktion der kanonischen Form der einzelnen deutschsprachigen Nomen und
4. Extraktion der kanonischen Formen der jeweiligen Synonyme.

Zur Extraktion wurde u.a. die folgende **SPARQL** Anfrage verwendet:

```
SELECT DISTINCT ?germannoun ?synonym WHERE {
 ?n a <http://www.lexinfo.net/ontology/2.0/lexinfo#Noun> .
 ?n <http://www.w3.org/ns/lemon/lime#language> "de" .
 ?n <http://kaiko.getalp.org/dbnary#synonym> ?p .
 ?p <http://kaiko.getalp.org/dbnary#describes> ?n2 .
 ?n2 a <http://www.lexinfo.net/ontology/2.0/lexinfo#Noun> .
 ?n <http://www.w3.org/ns/lemon/ontolex#canonicalForm> ?c .
 ?c <http://www.w3.org/ns/lemon/ontolex#writtenRep> ?germannoun .
 ?n2 <http://www.w3.org/ns/lemon/ontolex#canonicalForm> ?c2 .
 ?c2 <http://www.w3.org/ns/lemon/ontolex#writtenRep> ?synonym .
}
```

Code: SPARQL Anfrage zur Extraktion von deutschsprachigen Synonymen

Als Ergebnis wurden extrahiert:

- **6.668 deutschsprachige Nomen**, für die Synonyme bekannt sind und
- **21.634 Synonyme zu den entsprechenden Nomen**

Abschließend wurden die extrahierten Nomen mit den Titeln und Beschreibungestexten der Metadaten aus mCLOUD und GovData abgeglichen und die Anzahl der Nomen eingeschränkt. Als Ergebnis stehen bereit:

- **1.497 Nomen** aus **mCLOUD** und **GovData** sowie die entsprechenden Synonyme

3 Integration in Elasticsearch

Die Einbettung der generierten Synonymlisten in ElasticSearch kann über eine *Einfache Kontraktion* oder eine *Einfache Expansion* [ES-synonyms] sowie *Synonym Token Filter* [ES-tokenfilter] umgesetzt werden. Für jedes der rund 1.500 Nomen und deren entsprechenden Synonymlisten geschieht die Konfiguration zur Kontraktion über Zeilen der folgenden Form:

```
s-bahn, straßenbahn => stadtbahn
niederlassung => standort
```

Damit würde ein Suche nach “Straßenbahn an Niederlassung Paderborn” folgendermaßen erweitert:

- (straßenbahn,stadtbahn)
- (an)
- (niederlassung,standort)
- (paderborn)

Die Integration geschieht über die Elasticsearch Konfiguration und die Angabe einer Datei, die die Synonymlisten enthält. Dadurch ist eine dynamische Einbettung und Verknüpfung mit den OPAL Indexstrukturen in Elasticsearch möglich [OPAL-Index].

4 Zukünftige Arbeiten

Mögliche zukünftige Erweiterungen der Komponente bestehen im Wesentlichen aus den folgenden Arbeiten:

- Erweiterung der Synonymliste um Pluralformen
- Erweiterung der Synonymliste durch Analyse und Integration zusätzlicher Datensätze
- Erweiterung der Synonymliste um weitere Sprachen
- Einschränkung der Synonymliste durch Sichtung ungebräuchlicher Nomen

5 Literatur und weiterführende Verweise

- [BabelNet] **BabelNet**. https://babelnet.org/
- [DBnary] **DBnary**. http://kaiko.getalp.org/
- [DBnary15] Sérasset, Gilles: **DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF**. Semantic Web – Interoperability, Usability, Applicability, IOS Press, 2015, Multilingual Linked Open Data, 6 (4), pp.355-361. http://www.semantic-web-journal.net/system/files/swj648.pdf
- [Elasticsearch] **Elasticsearch**. https://www.elastic.co/products/elasticsearch
- [ES-synonyms] Gormley, Clinton; Tong, Zachary: **Synonyms. In: Elasticsearch: The Definitive Guide**. https://www.elastic.co/guide/en/elasticsearch/guide/current/synonyms.html
- [ES-tokenfilter] **Synonym Token Filter. In: Elasticsearch Reference.** https://www.elastic.co/guide/en/elasticsearch/reference/2.4/analysis-synonym-tokenfilter.html
- [lemon] **Lexicon model for ontologies (lemon)**. https://lemon-model.net/
- [lemonw3] Cimiano, Philipp; McCraeLexicon, John P.; Buitelaar, Paul: **Model for Ontologies**: Final Community Group Report 10 May 2016. https://www.w3.org/2016/05/ontolex/
- [LexInfo] **LexInfo.** https://www.lexinfo.net/
- [LingHub] **LingHub**. http://linghub.org/
- [OPAL-Index] Caglar Demir: OPAL D4.3 Prototype **index structures** and entity recognition
- [OPAL-LA] **OPAL linguistic-ambiguities**. https://github.com/projekt-opal/linguistic-ambiguities
- [Wiktionary] **Wiktionary**. https://www.wiktionary.org/
- [WordNet-RDF] RDF/OWL Representation of WordNet https://www.w3.org/TR/wordnet-rdf/